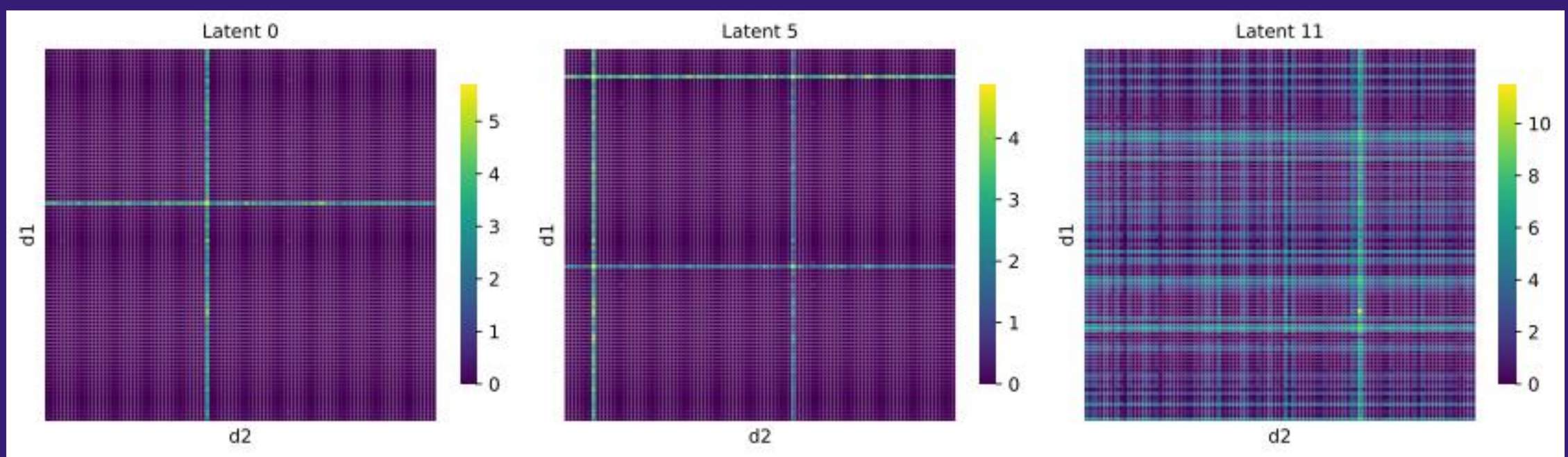
SAEs can reliably learn single latents for multiple features based on activation magnitude.



Sparse Autoencoders Can Learn Graded Latents for Relational Composition

Theo Farrell, Patrick Leask, Noura Al Moubayed

Durham University

1 TL;DR

- On a toy task, we find **SAEs reliably learn graded latents**, which are latents that correspond to different features depending on activation magnitude.
- This undermines some current SAE analysis which treats SAE latents as on/off switches for features.
- Future work should search for graded latents in pretrained LLM SAEs and move beyond the toy task.

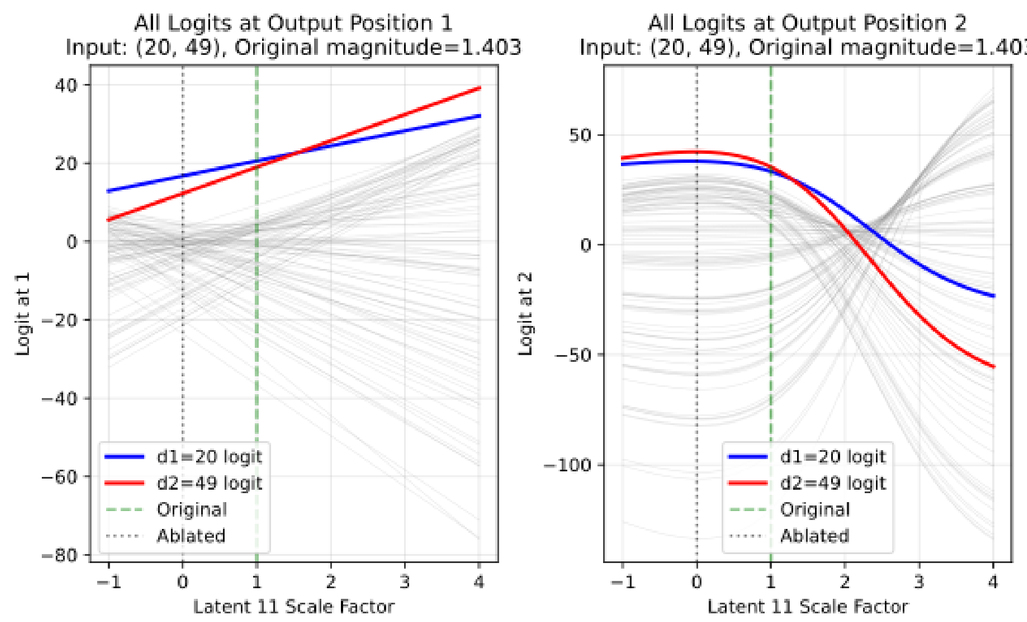
2 Background / Motivating Question

- SAE interpretability often assumes SAE latents are binary on/off switches for monosemantic features.
- However, transformers can encode different features (e.g. token ordering) by varying magnitudes along shared residual stream dimensions.
- This suggests SAEs should learn latents for each shared dimension but is not tested empirically.
- We provide empirical evidence of how SAEs respond to these features and attempt model analysis using them.

3 Finding Graded Latents

- We train **BatchTopK, JumpReLU and Matryoshka SAEs** on the residual stream of an attention-only transformer trained to copy bigrams via a single token: [a, b] → [s] → [a, b].
- Here, bigram order ([a, b] vs [b, a]) is encoded at [s] using the difference in attention from [s] to the inputs: $\Delta\alpha = \alpha_{s\to a} - \alpha_{s\to b}$.
- High-performing SAEs consistently learn two types of latents:
  - **k-symbol detectors** (e.g. latents 0, 5) – activate if an input bigram contains a symbol from a set of size k.
  - **Graded latents** (e.g. latent 11) – activate on almost all inputs with a magnitude strongly positively correlated with $\Delta\alpha$. This correlation is an empirically reliable identification method for graded latents.
- We steer on graded SAE latents by scaling their activation to observe the downstream effect.

4 Steering



- For one BatchTopK SAE, steering a graded latent swapped the output bigram order (e.g. from [a, b] to [b, a]) for nearly 50% of the dataset. See above figure for an example bigram.
- Non-graded latents swapped between 0-0.5%, demonstrating that graded latents are significantly different.
- This behaviour held across SAE architectures: high-performing SAEs consistently learned graded latents that could be steered to reliably swap outputs.

5 Limitations and Discussion

- Controlled toy setting – need to extend search to pretrained LLM SAEs such as Gemma Scope, for which graded latents could significantly affect analysis.
- Single-latent steering underestimates multi-latent effects. Co-activating latents can open additional swap ranges.
- In summary, this paper gives a controlled example where SAE latent activation values matter for interpretation.

PAPER

CODE

theodore.farrell@durham.ac.uk